Available online at www.sciencedirect.com
SCIENCE DIRECT

C. R. Acad. Sci. Paris, Ser. I 340 (2005) 319–324

COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Numerical Analysis/Partial Differential Equations

Numerical solution of the Monge–Ampère equation by a Newton's algorithm

Grégoire Loeper [a], Francesca Rapetti [b]

[a] Département de mathématiques, École polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland
[b] Laboratoire J.-A. Dieudonné, CNRS & université de Nice et Sophia-Antipolis, parc Valrose, 06108 Nice cedex 02, France

Received 16 November 2004; accepted after revision 17 December 2004

Available online 21 January 2005

Presented by Roland Glowinski

Abstract

We solve numerically the Monge–Ampère equation with periodic boundary condition using a Newton's algorithm. We prove convergence of the algorithm, and present some numerical examples, for which a good approximation is obtained in 10 iterations. ***To cite this article: G. Loeper, F. Rapetti, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

Résumé

Une méthode numérique de résolution de l'equation de Monge–Ampère. Nous résolvons numériquement l'équation de Monge–Ampère avec donnée au bord périodique en utilisant un algorithme de Newton. Nous prouvons la convergence de l'algorithme, et présentons quelques exemples numériques, pour lesquels une bonne approximation de la solution est obtenue en 10 itérations. ***Pour citer cet article : G. Loeper, F. Rapetti, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

Version française abrégée

Nous nous intéressons à la résolution numérique dans $\mathbb{R}^d$, $d \geqslant 2$, de l'équation de Monge–Ampère (1). Pour des fonctions $\psi : \mathbb{R}^d \mapsto \mathbb{R}$, convexes, l'Éq. (1) est de type elliptique non-linéaire. L'existence de solutions classiques pour cette équation se prouve par la méthode de continuité [7]. L'algorithme de Newton que nous adoptons pour résoudre (1) numériquement peut être considéré comme une mise en œuvre de cette méthode. Cette dernière s'appuie de manière essentielle sur les estimations a priori des dérivées secondes de la solution de (1), et nous nous appuyons également sur ces estimations pour prouver la convergence de l'algorithme (Théorème 2.1). Les

E-mail addresses: gregoire.loeper@epfl.ch (G. Loeper), frapetti@math.unice.fr (F. Rapetti).

1631-073X/$ – see front matter © 2004 Académie des sciences. Published by Elsevier SAS. All rights reserved.
doi:10.1016/j.crma.2004.12.018

experiences numériques ont été menées en dimension 2 et 3, mais les résultats théoriques restent valables en toute dimension. Du point de vue computationnel, à chaque itération de l'algorithme de Newton, les dérivées secondes de la fonction u sont approchées par un schéma de différences finies centré d'ordre 2 et le système (4) est résolu itérativement par une procedure BiCG non préconditionné. Les résultats numériques montrent la flexibilité et l'efficacité de l'algorithme en termes de nombre d'itérations et de temps de calcul pour une taille fixée du système (4). Une conclusion importante de ces travaux, est que l'on peut résoudre numériquement une équation elliptique pleinement non-linéaire au prix d'un nombre fini (i.e., indépendant de la taille de la grille) de problèmes elliptiques linéaires, au coût optimal $\mathrm{O}(N\log N)$ si l'on dispose d'un solveur multi-grille pour problèmes linéaires. Les limitations de la méthode sont la restriction à des densités suffisamment régulières (Hölder continues).

1. Introduction

We are here interested by the numerical solution of the Monge–Ampère equation

$$\det D^2\psi = \rho, \quad \psi \text{ convex over } \mathbb{R}^d,\ d \geqslant 2, \tag{1}$$

where $D^2\psi = (D_{ij}\psi)_{i,j=1,d}$, denotes the Hessian matrix of ψ and ρ is a given positive function. For a convex function ψ, Eq. (1) belongs to the class of fully non-linear elliptic equations. This class of equations has been a source of intense investigations in the last decades, with the theory of viscosity solutions [4]. Eq. (1) is also related to many areas of mathematics, such as geometry and optimal transportation (see [2,10] and the references therein). One of the crucial tools for proving the existence and regularity of a solution to this equation is the validity of a priori estimates on the solution's second order derivatives; these estimates allow to use the well known continuity method [7], in order to state the existence of (smooth) solutions. To obtain a solution of the Monge–Ampère equation, we implement a Newton's algorithm, which can be seen as a variant of the continuity method. The convergence of the algorithm is proved, for smooth enough right-hand sides, by using the same a priori estimates as before; these estimates allow to control the linearized problem, starting point of the algorithm formulation. The theoretical results we present are valid in any dimension, even if numerical experiments have been done in $\mathbb{R}^2$ and $\mathbb{R}^3$. We will be concerned here only with periodic boundary conditions in order to avoid, in a first time, problems arising from the boundary. In the periodic setting, Eq. (1) reads as follows: given a positive periodic function ρ on $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, find a periodic function $u : \mathbb{T}^d \to \mathbb{R}$ such that

$$F(u) := \det(I + D^2u) = \rho, \qquad \mathbf{x} \mapsto |\mathbf{x}|^2/2 + u \text{ convex over } \mathbb{R}^d. \tag{2}$$

Note that a necessary condition for Eq. (2) to be well-posed is that $\int_{\mathbb{T}^d}\rho = 1$. Wishing to solve (2) by using a Newton's algorithm, we need to linearize the operator F. Given A, B two $d \times d$ matrices, $\det(A + s\,B) = \det A + s\ \mathrm{trace}(A^t_{\mathrm{com}}B) + \mathrm{o}(s)$, where $s \in \mathbb{R}$ and A_{com} is the co-matrix of A, i.e., $A_{\mathrm{com}} = (\det A)\ A^{-1}$, provided A is invertible. This yields

$$F(u + s\,v) = \det\bigl(I + D^2(u + s\,v)\bigr) = \det(I + D^2u) + s\ \mathrm{trace}\bigl([I + D^2u]^t_{\mathrm{com}} D^2v\bigr) + \mathrm{o}(s),$$

for a smooth periodic function v and a parameter $s \in \mathbb{R}$. The linearized Monge–Ampère operator reads

$$DF(u)\cdot v = \sum_{i=1}^{d} M_{ij}\,D_{ij}v, \tag{3}$$

where $M = (M_{ij})_{i,j=1,d}$ is the co-matrix of $(I + D^2u)$. Eq. (2) being fully non-linear, we see that the coefficients of the linearized problem are second order derivatives of the solution itself, which explains the need for a priori estimates on these derivatives to control the linearized problem.

2. The algorithm: presentation and proof of convergence

The *algorithm* we consider to solve Eq. (3) reads: Given u_0, loop over $n \in \mathbb{N}$,

- *Computation of* $\rho_n = \det(I + D^2 u_n)$.
- *Assembling of* M^n *the co-matrix of* $(I + D^2 u_n)$.
- *Solution of the linearized Monge–Ampère equation*

$$\sum_{i,j=1}^{d} M_{ij}^n D_{ij}\theta_n = \frac{1}{\tau}(\rho - \rho_n). \tag{4}$$

- *Computation of* $u_{n+1} = u_n + \theta_n$.

The stabilization factor $\tau \geqslant 1$ is useful in the proof of the following convergence theorem

Theorem 2.1. *Let* ρ *be a positive probability density on* $\mathbb{T}^d$ *belonging to* $C^\alpha(\mathbb{T}^d)$ *for some* $\alpha \in (0,1)$. *There exists* $\tau \geqslant 1$ *depending on* $\{\min_{\mathbf{x}\in\mathbb{T}^d} \rho, \max_{\mathbf{x}\in\mathbb{T}^d} \rho, \|\rho\|_{C^\alpha(\mathbb{T}^d)}\}$, *such that if* $(u_n)_{n\in\mathbb{N}}$ *is the sequence constructed by the above algorithm, it converges in* $C^{2,\alpha'}$ *to the* (*unique up to a constant*) *solution* u *of* $\det(I + D^2 u) = \rho$, *for every* $0 < \alpha' < \alpha$.

We recall a result of existence of smooth solutions to Eq. (2) for Hölder continuous, positive right-hand sides. This result gives us the a priori bound needed to show the convergence of *algorithm* (4).

Theorem 2.2 (Caffarelli, [3]). *Let* ρ *be a probability density over* $\mathbb{T}^d$ *such that* $m \leqslant \rho \leqslant M$ *for some pair* $(m, M) > 0$. *Let* $u : \mathbb{T}^d \mapsto \mathbb{R}$ *be solution of* $\det(I + D^2 u) = \rho$, *with* $u + |\cdot|^2/2$ *convex. Then there exists a non-decreasing function* $\mathcal{H}_{m,M}$ *such that* $\|u\|_{C^{2,\alpha}(\mathbb{T}^d)} \leqslant \mathcal{H}_{m,M}(\|\rho\|_{C^\alpha(\mathbb{T}^d)})$.

Proof of Theorem 2.1. The C^α norm $\|f\|_{C^\alpha(\mathbb{T}^d)}$ of a function f is defined by $\|f\|_{L^\infty(\mathbb{T}^d)} + \sup_{\mathbf{x},\mathbf{y}\in\mathbb{T}^d} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^\alpha}$. We prove the following *bounds* by induction: There exist $C_1 > 0$, $C_2 > 0$ depending on the quantities stated in Theorem 2.1 such that (i) $\frac{1}{C_1}\rho \leqslant \rho_n \leqslant C_1 \rho$ and (ii) $\|\rho - \rho_n\|_{C^\alpha} \leqslant C_2$.

For a smooth ρ_0 (note that, in practice, we shall take $u_0 = 0$, $\rho_0 = 1$) we can always find C_1, C_2 so that (i) et (ii) are satisfied. We suppose that (i) and (ii) hold true for ρ_n and show that they extend to ρ_{n+1}. We recall that θ_n is defined in (4) by $D\det(I + D^2 u_n) \cdot D^2\theta_n = M_{ij}^n D_{ij}\theta_n = \frac{1}{\tau}(\rho - \rho_n)$. We then have $\rho_{n+1} = \det(I + D^2 u_n + D^2\theta_n) = \det(I + D^2 u_n) + D\det(I + D^2 u_n) \cdot D^2\theta_n + r_n = \rho_n + \frac{1}{\tau}(\rho - \rho_n) + r_n$. Let us evaluate r_n: it consists of products of at least two second derivatives of θ_n and eventually second derivatives of u_n, depending on the dimension. Assuming that the bounds (i) and (ii) hold, Theorem 2.2 implies that $I + D^2 u_n$ and therefore M^n are C^α smooth, uniformly elliptic matrices. Since θ_n solves (4), from standard Schauder elliptic theory [7] we get that $\|D^2\theta_n\|_{C^\alpha} \leqslant \frac{C_3(C_1,C_2)}{\tau}\|\rho - \rho_n\|_{C^\alpha}$. Therefore

$$\|r_n\|_{C^\alpha} \leqslant C_4(C_1, C_2)\|\rho - \rho_n\|_{C^\alpha}^2 \frac{1}{\tau^2}. \tag{5}$$

Combining with the identity

$$(\rho - \rho_{n+1})(\mathbf{x}) = \left(1 - \frac{1}{\tau}\right)(\rho - \rho_n)(\mathbf{x}) + r_n(\mathbf{x}), \tag{6}$$

we obtain

$$\|\rho - \rho_{n+1}\|_{C^\alpha} \leqslant \left(1 - \frac{1}{\tau}\right)\|\rho - \rho_n\|_{C^\alpha} + \frac{C_4}{\tau^2}\|\rho - \rho_n\|_{C^\alpha}^2. \tag{7}$$

By the induction assumption (ii), $\|\rho - \rho_n\|_{C^\alpha}$ is bounded by C_2, and the inequality (7) implies that $\|\rho - \rho_{n+1}\|_{C^\alpha} \leqslant \|\rho - \rho_n\|_{C^\alpha}(1 - \frac{1}{\tau} + \frac{C_4 C_2}{\tau^2})$. This is smaller than $\|\rho - \rho_0\|_{C^\alpha}$ if $\frac{C_4 C_2}{\tau} \leqslant 1$, thus for τ large enough depending on C_1, C_2. So far we have checked that the bound (ii) is preserved for τ large enough.

Let us now check bound (i): Let $m=\inf_{\mathbf{x}\in\mathbb{T}^d}\rho(\mathbf{x})$, $M=\sup_{\mathbf{x}\in\mathbb{T}^d}\rho(\mathbf{x})$ (we recall that $m>0$). The induction assumption (i) says that $(\rho-\rho_n)(\mathbf{x})\leqslant\rho(\mathbf{x})(1-1/C_1)$. Then (5) implies $\|r_n\|_{L^\infty}\leqslant\frac{C_5(C_1,C_2)}{\tau^2}$ and this bound combined with (6) yields $(\rho-\rho_{n+1})(\mathbf{x})\leqslant\frac{\tau-1}{\tau}(\rho-\rho_n)(\mathbf{x})+\frac{C_5}{\tau^2}\leqslant\frac{\tau-1}{\tau}\rho(\mathbf{x})(1-1/C_1)+\frac{C_5}{\tau^2}$. The last expression is smaller than $\rho(\mathbf{x})(1-1/C_1)$ for $\tau>\frac{C_5}{\rho(\mathbf{x})(1-1/C_1)}$. Therefore we conclude the following: if $\tau>\frac{C_5}{m(1-1/C_1)}$, bounds (i) and (ii) imply that $\rho_{n+1}\geqslant\rho/C_1$.

Now we follow the same strategy and use that $(\rho_n-\rho)(\mathbf{x})\leqslant(C_1-1)\rho(\mathbf{x})$ (still from bound (i)). We then check that for $\tau\geqslant\frac{C_5}{m(C_1-1)}$, we have also $(\rho_{n+1}-\rho)(\mathbf{x})\leqslant(C_1-1)\rho(\mathbf{x})$.

We conclude that for a choice of ρ_0 and $C_1>1, C_2>0$ that satisfy (i), (ii), there exists τ that depends only on $\{m,M,C_1,C_2\}$ such that bounds (i) and (ii) are preserved for all $n\in\mathbb{N}$.

Concerning the *convergence* of *algorithm* (4), from (7), we see that if $\|\rho-\rho_n\|_{C^\alpha}\leqslant\tau/(2C_4)$, we have a geometric convergence with rate at least $1-1/(2\tau)$. This will be satisfied for $\tau\geqslant 2C_2C_4$. Therefore ρ_n converges to ρ in C^α. From Theorem 2.2, the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $C^{2,\alpha}$; note also that we have imposed $u_n(\mathbf{0})=0$. Therefore by the Ascoli–Arzela's theorem, $(u_n)_{n\in\mathbb{N}}$ is precompact in $C^{2,\alpha'}$ for every $\alpha'<\alpha$. The solution of (2) being unique once we impose $u(\mathbf{0})=0$, the whole sequence must be converging to the solution u of (2). This ends the proof of Theorem 2.1. □

3. Numerical experiments

The computational domain for the *algorithm* is $\mathbb{T}^d$ which is reproduced by considering $V=[0,1]^d$ together with periodic boundary conditions. The solution of the linearized Monge–Ampère equation in V is unique, up to a constant that can be easily fixed by assigning the value of u at a given point of V. At each iteration n of the *algorithm*, the two matrices D^2u_n and M^n are assembled by means of a centered second order finite difference scheme on a Cartesian grid of N^d points over V. This means, e.g., that $(D_{12}u)_{i,j}\approx(u_{i+1,j+1}-u_{i-1,j+1}-u_{i+1,j-1}+u_{i-1,j-1})/(4h^2)$, with $u_{i,j}\approx u\ (i\,h,\,j\,h)$, $1\leqslant i,j\leqslant N$, $h=1/N$, and periodicity N in considering the indexes $i\pm 1$, $j\pm 1$. System (4) is then solved iteratively by a BiCG procedure [9], with stopping threshold on the residual norm equal to 10^{-8}. The BiCG algorithm is not preconditioned, and the average number of BiCG iterations to converge at each Newton's one goes from 30 on the coarsest grid up to 1000 on the finest. For the numerical tests, we consider a starting density $\bar\rho=1$ and a target density ρ of the form $\rho(\mathbf{x})=1+\beta\sin(2\pi kx)\sin(2\pi ky)$, with $0<\beta<1$ and $k\geqslant 1$. All shown results are obtained in $\mathbb{T}^d$, $d=2$; those for $d=3$ are similar.

Concerning the performances of the considered *algorithm*, Fig. 1(left) shows the convergence history of the error $\|\rho-\rho_n\|_{L^2(\Omega)}$. Different grids are used, from a coarse one, $N=16$, to a fine one, $N=512$, and in all the

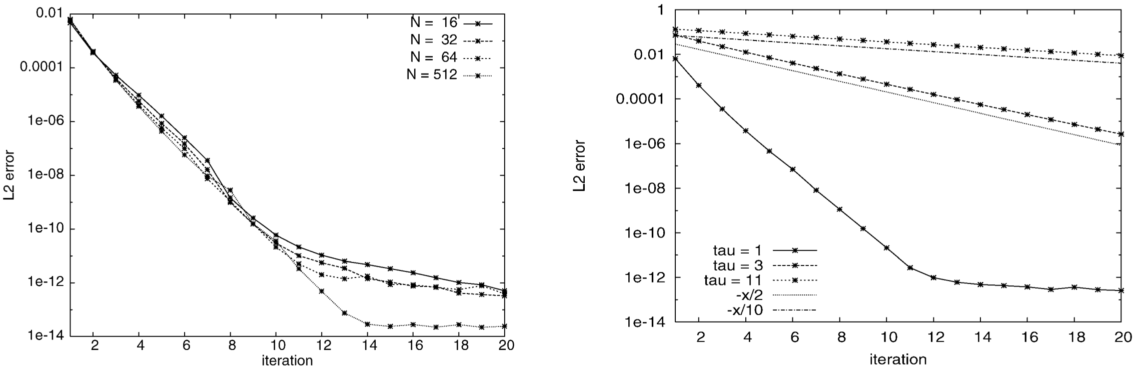

Fig. 1. Convergence history of the error on ρ ($\beta=0.8$ and $k=2$) in the L^2-norm over 20 iterations. A semi-logarithmic scale is used. Left: the linearized Monge–Ampère equation is solved on different grids with $\tau=1$. Right: the linearized Monge–Ampère equation is solved on a given grid ($N=64$) and for different values of τ.

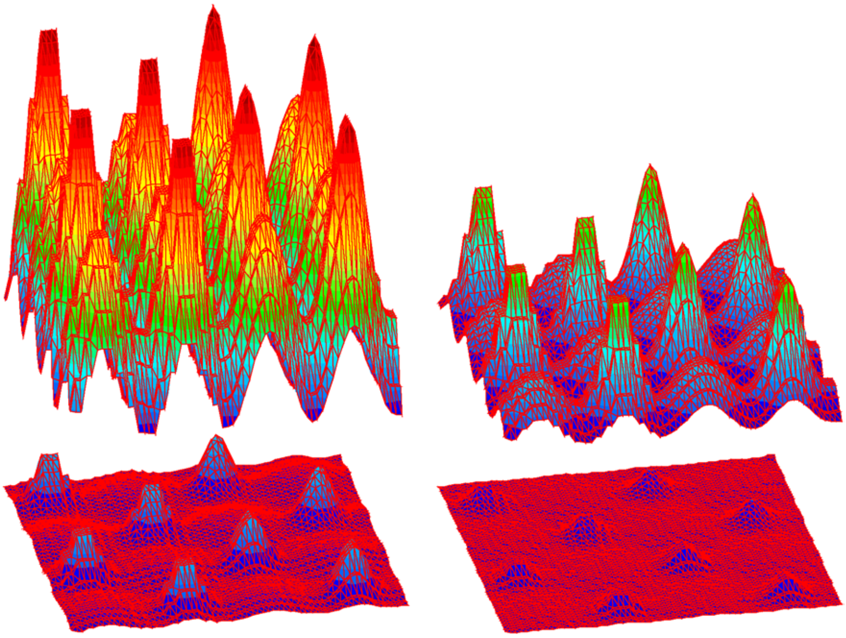

Fig. 2. Distribution over V of the error on ρ ($\beta = 0.8$, $k = 2$) with $N = 64$. The highest absolute value is 0.0609 ($n = 1$, top left), 0.0239 ($n = 2$, top right), 0.00949 ($n = 3$, bottom left) and 0.00379 ($n = 4$, bottom right), respectively.

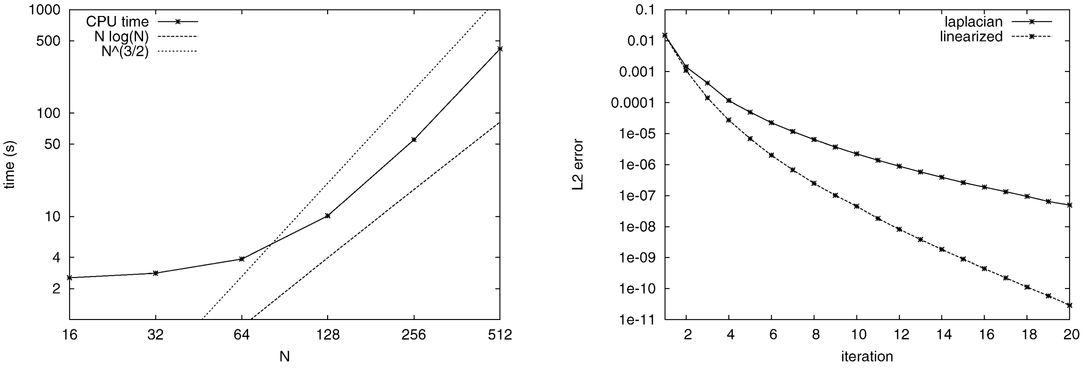

Fig. 3. (Left) CPU time in logarithmic scale for the *algorithm* with respect to N ($\beta = 0.8$, $k = 2$ and $\tau = 1$). (Right) Convergence history of the error on ρ ($\beta = 0.99$ and $k = 1$) in the L^2-norm over 20 iterations. The Laplace and the linearized Monge–Ampère equations are solved with $N = 128$ and $\tau = 1$.

cases, 10 Newton's iterations are enough to have an error $\approx 10^{-10}$. Note that in practice we have taken $\tau = 1$, and the convergence is faster than geometric. In Fig. 1(right) the convergence history of the error is shown together with the asymptotic behavior for three different values of τ.

The *algorithm* is quite flexible and efficient: similar results can be obtained on very coarse ($N = 16$) as well as on very fine ($N = 512$) grids to approximate a sine function. A variety of parameters β, k and τ can be selected. Moreover, the *algorithm* convergence is quite fast. In Fig. 2 are shown the distributions of the error on ρ at the grid points for the first 4 iterations. For the considered case, the highest absolute value of the error is reduced, in 4 iterations, to $O(10^{-3})$, with a dumping factor ≈ 2 at each iteration, in agreement with the convergence order of a Newton's algorithm.

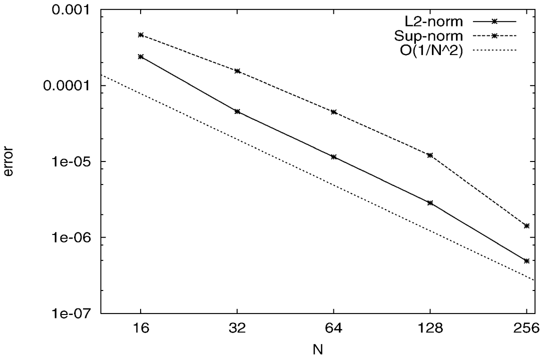

Fig. 4. Approximation error in the L^2 and L^∞ norms for the given function $u_{\mathrm{ex}}(\mathbf{x}) = \beta\sin(2\pi kx)\sin(2\pi ky)$, $\beta = 0.02$ and $k = 1$. The error values, presented in logarithmic scale, are obtained after 10 Newton's iterations for $N = 16, 32, 64, 128, 256$.

The CPU time curve for the considered *algorithm* is presented in Fig. 3(left). Note that this curve is in between the (optimal) $N\log(N)$ and the (asymptotic) $N^{3/2}$ ones.

A simplified version of the *algorithm* can be obtained by replacing the solution θ_n of the linearized Monge–Ampère equation (4) with that of the Laplace equation $\Delta\theta_n = \frac{1}{\tau}(\rho - \rho_n)$. This amounts to replace the co-matrix M^n with the identity matrix, for all $n \in \mathbb{N}$. Finite differences are thus involved to discretize the Laplace operator, which has smoother coefficients, and guarantees strict ellipticity. For a smooth right-hand side far from zero, the two methods do not differ too much, however, for a density that goes very close to 0, such as in the considered case with $\beta = 0.99$ and $k = 1$, the method based on the linearized Monge–Ampère equation gives better results, as it is shown in Fig. 3(right).

In Fig. 4 we present the behavior with respect to N of the errors $\|u - u_{\mathrm{ex}}\|_{L^2(V)}$ and $\|u - u_{\mathrm{ex}}\|_{L^\infty(V)}$ for $u_{\mathrm{ex}}(\mathbf{x}) = \beta\sin(2\pi kx)\sin(2\pi ky)$, $\beta = 0.02$, $k = 1$. The asymptotic order in both cases is $\mathrm{O}(\frac{1}{N^2})$.

We have presented an efficient algorithm to solve the Monge–Ampère equation for smooth right-hand sides. In this case, the cost of the algorithm is very close to optimal, since the convergence is obtained in $\mathrm{O}(N^{3/2})$ seconds, with a finite number of Newton's iterations, each one being the approximation of a linear elliptic problem. We see that solving a fully non-linear elliptic equation can be done at the cost of solving a finite number of linear elliptic problems. The convergence of the method is not guaranteed for non-smooth right-hand side and alternative approaches are proposed in [1,5,6,8].

References

[1] J.-D. Benamou, Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem, Numer. Math. 84 (3) (2000) 375–393.

[2] Y. Brenier, U. Frisch, M. Henon, G. Loeper, S. Matarrese, R. Mohayaee, A. Sobolevskiĭ, Reconstruction of the early Universe as a convex optimization problem, Mon. Not. R. Astron. Soc. 346 (2) (2003) 501–524.

[3] L. Caffarelli, Interior $W^{2,p}$ estimates for solutions of Monge–Ampère equation, Ann. Math. (2) 131 (1) (1990) 135–150.

[4] L. Caffarelli, X. Cabre, Fully Nonlinear Elliptic Equations, Amer. Math. Soc. Coll. Publ., vol. 43, American Mathematical Society, Providence, RI, 1995.

[5] E. Dean, R. Glowinski, Numerical solution of the two-dimensional elliptic Monge–Ampère equation with Dirichlet boundary conditions: an augmented Lagrangian approach, C. R. Acad. Sci. Paris, Ser. I 336 (9) (2003) 779–784.

[6] E. Dean, R. Glowinski, Numerical solution of the two-dimensional elliptic Monge–Ampère equation with Dirichlet boundary conditions: a least square approach, C. R. Acad. Sci. Paris, Ser. I 339 (12) (2004) 887–892.

[7] D. Gilbarg, N. Trudinger, Elliptic Partial Differential Equations of Second Order, second ed., Grundlehren Math. Wiss. [Fund. Princ. Math. Sci.], vol. 224, Springer-Verlag, Berlin, 1983.

[8] V.I. Ollicker, L.D. Prussner, On the numerical solution of the equation $z_{xx}z_{yy} - z_{xy}^2 = f$ and its discretization. I, Numer. Math. 54 (1988) 271–293.

[9] Q. Quarteroni, A. Valli, Numerical Approximation of Partial Differential Equations, Comput. Math., vol. 23, Springer-Verlag, Berlin, 1994.

[10] C. Villani, Topics in Optimal Transportation, Graduate Ser. in Math., American Mathematical Society, 2003.